arXiv:2503.13401v1 [cs.CL] 17 Mar 2025

Using the Tools of Cognitive Science to Understand Large Language Models at Different Levels of Analysis

Alexander Ku[1,2], Declan Campbell[2], Xuechunzi Bai[3], Jiayi Geng[4], Ryan Liu[4], Raja Marjieh[1], R. Thomas McCoy[5], Andrew Nam[6], Ilia Sucholutsky[7], Veniamin Veselovsky[4], Liyi Zhang[4], Jian-Qiao Zhu[4], and and Thomas L. Griffiths[1,4,6]

[1]Department of Psychology, Princeton University

[2]Princeton Neuroscience Institute, Princeton University

[3]Department of Psychology, The University of Chicago

[4]Department of Computer Science, Princeton University

[5]Department of Linguistics, Yale University

[6]Princeton Laboratory for Artificial Intelligence, Princeton University

[7]Center for Data Science, New York University

Abstract

Modern artificial intelligence systems, such as large language models, are increasingly powerful but also increasingly hard to understand. Recognizing this problem as analogous to the historical difficulties in understanding the human mind, we argue that methods developed in cognitive science can be useful for understanding large language models. We propose a framework for applying these methods based on Marr's three levels of analysis. By revisiting established cognitive science techniques relevant to each level and illustrating their potential to yield insights into the behavior and internal organization of large language models, we aim to provide a toolkit for making sense of these new kinds of minds.

Keywords: large language models, levels of analysis

Using the Tools of Cognitive Science to Understand Large Language Models at Different Levels of Analysis

Introduction

The last decade has seen a series of breakthroughs in artificial intelligence (AI) research, culminating in the creation of the large language models that underlie chat-based agents such as ChatGPT, Claude, Gemini, and LLaMa (Achiam et al., 2023; Team et al., 2023; Touvron et al., 2023). These breakthroughs have been driven by a specific strategy: starting with generic artificial neural network architectures and increasing their size and training data. Artificial neural networks are notoriously difficult to interpret, finding solutions to problems that are expressed in the form of billions of continuous weighted connections between units. As a consequence, computer scientists now face an unfamiliar problem: they have created systems that they do not understand. To make things even worse, since the training data and weights of many of the leading systems are not available outside the companies that created them, in many cases the only insights we can obtain about the nature of these systems are those that can be gleaned by studying their behavior.

Even though this problem is unfamilar to computer scientists, it is very familiar to another group of researchers: cognitive scientists. Cognitive science is the interdisciplinary science of the mind, and for the 70 or so years since its inception (Miller, 2003) has been limited by the fact that it had relatively few kinds of mind to study. To cognitive scientists, the advent of intelligent machines offers exciting new opportunities to apply methods that have been refined through trying to understand how human minds work (Binz and Schulz, 2023; Coda-Forno et al., 2024). Those methods encompass both techniques that are based on human behavior and insights that come from related fields such as neuroscience that explore the underlying mechanisms.

In this paper we summarize some of the tools from cognitive science that we believe are particularly useful for understanding large language models (and related approaches such as vision-language models). We revisit some of the “greatest hits” of cognitive science

– techniques such as rational analysis, the axiomatic approach, and multidimensional scaling – and illustrate how they translate into techniques for gaining insight into AI systems. We illustrate the potential of this approach by using examples from our own work and by drawing analogies between nascent methods in computer science and tools developed in other fields.

As an organizing principle, we divide these tools up by the different kinds of questions to which they provide answers. The computational neuroscientist David Marr (Marr, 1982) suggested that information processing systems can be understood at three levels of analysis: the *computational* level, which focuses on the abstract computational problem a system solves and its ideal solution; the *algorithmic* level, which focuses on the representations and algorithms that approximate that solution; and the *implementation* level, which focuses on how those representations and algorithms are realized in a physical system. The same three levels can be used for analyzing large language models, focusing on how such systems are shaped by their function, the solutions that they seem to find, and the realization of those solutions in weights and units within the underlying artificial neural network (see Table 1). Drawing these parallels also provides a guide for where we might expect to look for relevant tools, with the computational level focusing on computational modeling techniques, the algorithmic level drawing on methods from cognitive psychology, and the implementation level being inspired by neuroscience. As a result, we see thinking at these different levels of analysis as being particularly productive for making sense of large language models.

Computational level: Considering function

The computational level focuses on the abstract problem that a system needs to solve. By recognizing the pressures that this problem exerts upon the system, we can make predictions about the properties that the system is likely to have (Shepard, 1987; Marr, 1982; Anderson, 1990; Griffiths, 2020). This perspective is perhaps most familiar from evolutionary biology, in which organisms are understood through the lens of the

Table 1

Understanding natural and artificial minds across Marr's levels of analysis.

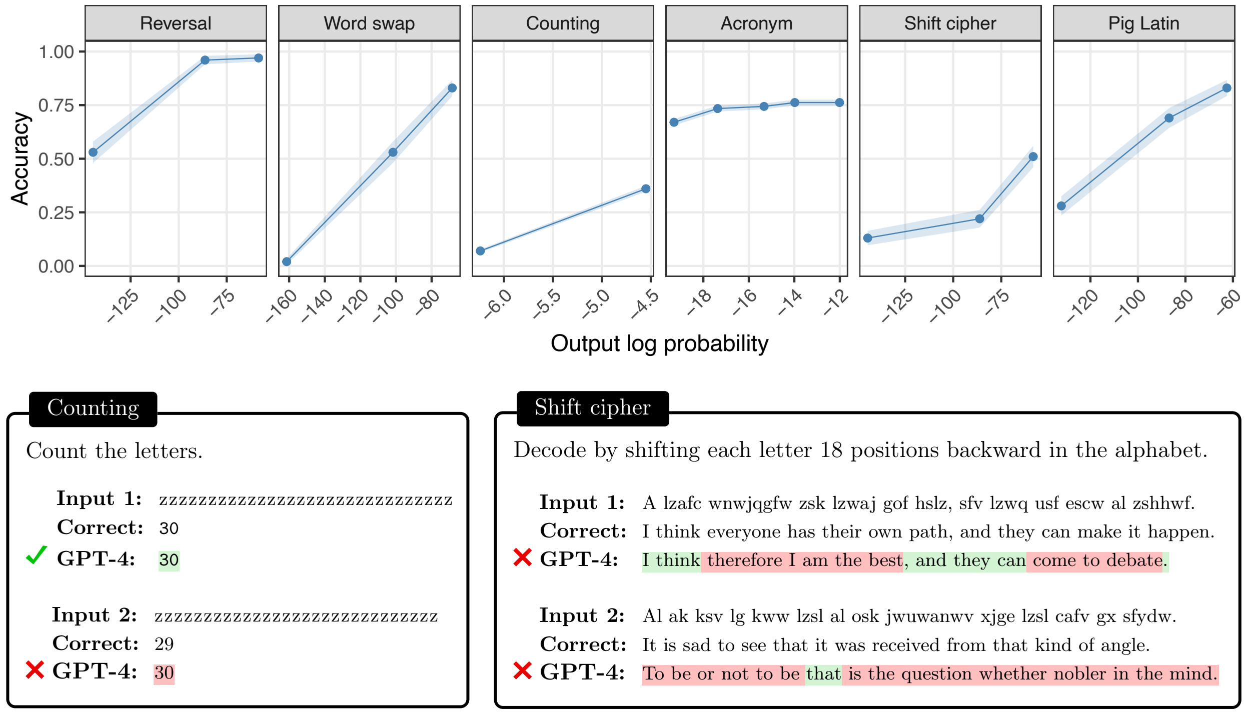
Level of analysis	Focus	Cognitive science	Artificial intelligence
Computational	What problem is being solved? What is the ideal solution?	Understanding behavior in terms of optimal solutions to environmental pressures (e.g., Bayesian models of cognition).	Understanding behavior in terms of the training objective (e.g., next-word prediction) and deviations from optimal benchmarks (e.g., violations of probability axioms).
Algorithmic	How is the solution approximated? What representations and processes are used?	Inferring mental representations and processes through behavioral experiments (e.g., reaction times, error patterns, similarity judgments).	Probing representations and processes via behavioral analysis (e.g., analyzing systematic errors, soliciting similarity judgments, uncovering hidden associations).
Implementation	How are the representations and algorithms physically realized?	Studying neural circuits and population activity using techniques like optogenetics, single-unit recording, fMRI, and MVPA.	Studying artificial neurons, circuits (e.g., induction heads), and population activity using techniques like activation patching, sparse autoencoders, and representational geometry analysis.

evolutionary pressures that have shaped them. For instance, our understanding of bird flight is informed by the aerodynamic principles that bird flight must obey. Similarly, we can gain insight into intelligent systems by considering how they have been shaped by the functions that they must perform.

This emphasis on function makes the computational level well-suited for analyzing AI systems. Although many aspects of modern AI systems are difficult to interpret (including their behavior and the mechanisms that they use to achieve that behavior), one aspect that we understand well is the function that the system is optimized to perform. Specifically, this function is explicitly defined by humans in the form of the AI system's training objective. Therefore, computational-level analysis enables us to start from something we understand well – the training objective – and use it to reason about the less-well-understood territory of the behavior of the resulting systems.

The embers of autoregression

McCoy et al. (McCoy et al., 2024) used an approach based on computational-level analysis to try to understand the behavior of large language models (LLMs). For these



Counting

Count the letters.

Input 1: zzzzzzzzzzzzzzzzzzzzzzzzzzzzzz
Correct: 30
✓ **GPT-4:** 30

Input 2: zzzzzzzzzzzzzzzzzzzzzzzzzzzzz
Correct: 29
✗ **GPT-4:** 30

Shift cipher

Decode by shifting each letter 18 positions backward in the alphabet.

Input 1: A lzafc wnwjqgfw zsk lzwaj gof hslz, sfv lzwq usf escw al zshhwf.
Correct: I think everyone has their own path, and they can make it happen.
✗ **GPT-4:** I think therefore I am the best, and they can come to debate.

Input 2: Al ak ksv lg kww lzsl al osk jwuwanwv xjge lzsl cafv gx sfydw.
Correct: It is sad to see that it was received from that kind of angle.
✗ **GPT-4:** To be or not to be that is the question whether nobler in the mind.

Figure 1

Large language models perform better when they need to produce a high-probability piece of text than when they need to produce a low-probability piece of text, even in deterministic settings where probability should not matter.

systems, the primary training objective is next-token prediction, also known as autoregression: predicting the next token (word or part of a word) in a piece of text given the preceding tokens. Analysis of this task leads to the prediction that LLMs will perform better when they need to produce a high-probability piece of text than when they need to produce a low-probability piece of text, even in deterministic settings where probability should not matter. Intuitively, this prediction follows from the way in which next-token prediction fundamentally depends on the probabilities of token sequences; this intuition is derived more formally in (McCoy et al., 2024) via a Bayesian analysis of autoregression.

The prediction that LLM behavior will be sensitive to probability is borne out in experiments testing a range of LLMs across a range of tasks (see Figure 1). For instance, when GPT-4 is asked to count how many letters there are in a list, it performs much better when the answer is a frequently-used number than a more rarely-used number; e.g., when the answer is 30, its accuracy is 97%, but when the answer is 29, its accuracy is only 17%, presumably because the number 30 is used more often in natural text than the number 29. Similarly, when asked to decode a message written in a simple cipher, GPT-4's accuracy was 51% when the answer was a high-probability sentence but only 13% when the answer was a low-probability word sequence. Thus, even though LLMs can be applied to many different tasks – a capability that has been viewed as evidence that LLMs show "sparks of artificial general intelligence" (Bubeck et al., 2023) – they also continue to display "embers of autoregression" – behavioral trends that reflect the nature of the specific objective they were optimized for.

Bayesian optimality as a benchmark

The "embers of autoregression" example illustrates a broader principle: many cognitive tasks can be characterized as inductive inference problems under uncertainty (Griffiths et al., 2024a). Bayesian models of cognition provide optimal solutions to these problems, and have become instrumental in explaining human performance across diverse domains, including perception (Yuille and Kersten, 2006), language processing (Chater and Manning, 2006; Griffiths et al., 2007), categorization (Sanborn et al., 2010), and intuitive physics (Sanborn et al., 2013; Battaglia et al., 2013). We can compare LLMs to these optimal solutions to gain insight into their behavior, much as cognitive scientists use Bayesian models to understand human cognition.

The connection between Bayesian optimality and LLMs can be made more explicit by considering the problem of next-token prediction that is typically used in training these models. Predicting the next token can be done by extracting the predictive sufficient statistics from previous tokens (Bernardo and Smith, 1994). For some datasets, the Bayesian posterior distribution over particular parameters or hypotheses about the generating process can serve as a predictive sufficient statistic (Zhang et al., 2024). This perspective can be used to understand the representations that LLMs form and how they should relate to ideal Bayesian solutions (Zhang et al., 2023, 2024). Several other recent papers have also identified interesting connections between LLMs and Bayesian inference (Xie et al., 2021; Wang et al., 2024a; Zheng et al., 2023).

These connections suggest that we might be able to create simple Bayesian models of the inferences drawn by LLMs. Such Bayesian models can be used to explore the implicit prior distributions adopted by LLMs and to compare the resulting distributions with those inferred from human behavior. For example, Griffiths et al. (Griffiths et al., 2024b) used a simple Bayesian model of predicting the future (Griffiths and Tenenbaum, 2006) to recover implicit prior distributions about the extent or duration of phenomena from GPT-4. Zhu and Griffiths (Zhu and Griffiths, 2024b) built on this work, using an iterated learning procedure originally developed for sampling from human priors to explore the priors that LLMs have for causal relationships and probability distributions, as well as their implicit assumptions about speculative events such as the development of superhuman AI.

Violations of axiomatic systems

Another approach to understanding intelligent systems at the computational level is to consider how their behavior relates to an axiomatic system that captures the solution to a problem. One of the original examples used by Marr had this flavor: he suggested that we can understand the computational problem solved by a cash register by recognizing that the expectations we have about shopping, such as the fact that the order in which items are checked out doesn't affect the total price, correspond to the axiomatic system of arithmetic. In cognitive science, the most celebrated application of this approach has been decision theory. By considering how to define rationality, decision theorists were able to specify a set of axioms that result in the discovery that rational agents should seek to maximize expected utility (Von Neumann and Morgenstern, 1947; Savage, 1954). Asking whether this axiomatic system actually describes human behavior resulted in fundamental insights into human decision-making, with Kahneman and Tversky carrying out an

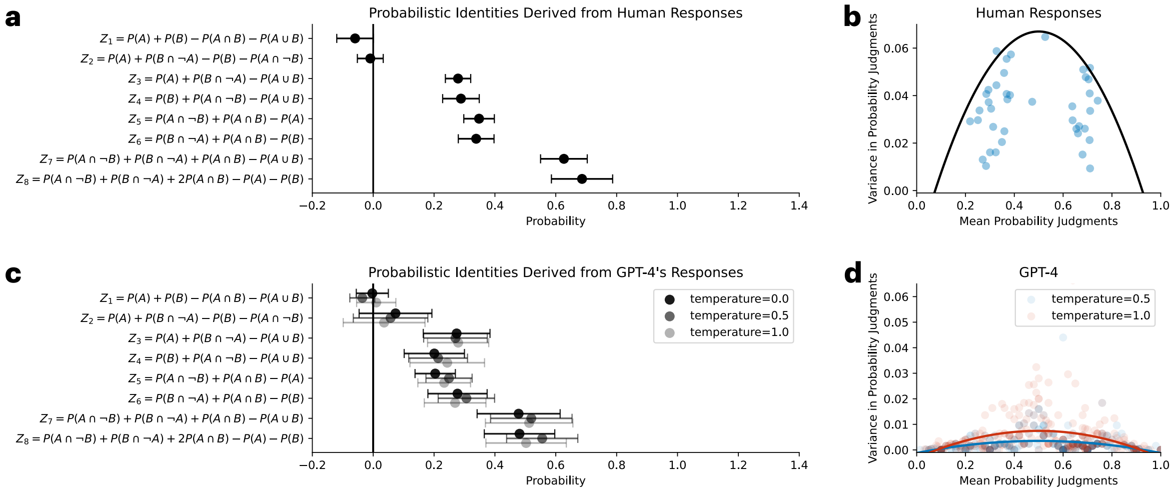


Figure 2

*Incoherent probability judgments from humans (**a**, **b**) and GPT-4 (**c**, **d**). Like human probability judgments (**a**), GPT-4's judgments systematically deviate from zero when combined into probabilistic identities (**c**). When repeatedly queried about the same event, the mean-variance relationship of probability judgments follows an inverted-U shape for both humans (**b**) and GPT-4 (**d**). Human data are adapted from (Zhu et al., 2020), GPT-4 results are from (Zhu and Griffiths, 2024a).*

influential research program that showed that people systemically violate the prescriptions of these axioms (Tversky and Kahneman, 1974; Kahneman and Tversky, 1979).

Considering relevant axiomatic systems – and discovering how they are violated – provides another tool for understanding LLMs. For example, probability theory dictates that the probabilities of an event A and its complement $\neg A$ sum to 1, meaning $P(A) + P(\neg A) = 1$. To assess whether LLMs adhere to this rule, we can examine deviations from zero in $P(A) + P(\neg A) - 1$. Similarly, other probabilistic identities can be tested, such as $P(A) + P(B) - P(A \wedge B) - P(A \vee B)$, which should also equal zero if judgments are coherent (Zhu et al., 2020). Eliciting probability judgments for logically related events from GPT and LLaMa models (see Figure 2), shows that probability identities formed using judgments generated by LLMs systematically deviate from zero, violating the rules of probability theory (Zhu and Griffiths, 2024a). In addition to maintaining coherence across logically related events, a rational agent should produce consistent probability judgments when repeatedly queried about the same event. However, repeated probability judgments elicited from LLMs exhibit an inverted-U-shaped mean-variance relationship (Zhu and Griffiths, 2024a). These systematic deviations also qualitatively mirror those observed in human probability judgments, suggesting that LLMs exhibit similar biases in probabilistic reasoning.

Optimal behavior can also be defined with respect to problems or situations with inherent uncertainty. For example, in a risky choice task where participants select one of many gambles (each gamble corresponding to a probabilistic distribution of outcomes), there is always a rational choice that maximizes the expected value (making the simplest possible assumption about the utility associated with an option by equating its utility with its monetary value). Here, using chain-of-thought reasoning (Wei et al., 2022) make choices almost completely rationally, but without such reasoning, their choices are noisy and sometimes ignore probabilities completely (Liu et al., 2024a). Furthermore, when LLMs are asked to predict human performance on the task, they predict humans to behave highly rationally, even though people behave much less so.

Summary

The computational level of analysis provides a powerful lens for understanding large language models by focusing on their function. Examining the training objective (in this case, autoregression) can directly predict specific behavioral patterns, as illustrated by the "embers of autoregression." Furthermore, comparing LLMs to optimal benchmarks, whether derived from Bayesian models of cognition or the axioms of probability and decision theory, can reveal both the surprising capabilities and the systematic limitations of these systems. By considering what computational problem LLMs are solving (or approximating), we can gain significant insight into their behavior and internal structure, even when the algorithmic and implementation details remain opaque.

Algorithmic level: Identifying representations and processes

Just as the computational level asks what problem an information-processing system solves, the algorithmic level explores how that solution is approximated. This level concerns itself with the specific representations and algorithms used to carry out the computation. Consider bird flight again: while aerodynamics dictates the principles of flight (lift, drag, thrust), different bird species employ different algorithms – variations in flapping patterns and soaring techniques – to achieve flight. These variations represent different algorithmic solutions to the same computational problem. In cognitive science, understanding the algorithmic level involves designing experiments that probe the internal workings of the mind, inferring the nature of mental representations and the processes that operate on them. Just as ornithologists might study wing movements and muscle activity to understand bird flight, cognitive psychologists use reaction times, error patterns, and carefully designed stimuli to understand human cognition. We can apply similar techniques to investigate the algorithmic solutions of large language models.

The algorithmic level is particularly relevant to LLMs because, unlike traditional symbolic AI systems with explicitly programmed rules, the specific algorithms and representations employed by LLMs are not pre-defined. They are learned through the training process, resulting in complex and often opaque internal structures that aren't obviously localized in any particular location in the network. Proprietary closed-source networks present additional challenges as their internal states are not directly observable. This problem is much the same as that faced by psychologists before the advent of modern neuroscientific tools and methodologies.

Cognitive psychology offers a rich toolbox of methods for exploring the algorithmic level, many of which can be creatively adapted to study LLMs. This section explores how cognitive science-inspired approaches – such as analyzing systematic error patterns, soliciting similarity judgments, and exploring associations – can be used to uncover the algorithms and representations of LLMs.

Parallel and serial processing

A fundamental challenge for any cognitive system, whether biological or artificial, is the tradeoff between processing information serially (one item at a time) or in parallel (multiple items simultaneously) (Treisman and Gelade, 1980; Townsend, 1990). Parallel processing offers efficiency, allowing for rapid processing of multiple inputs. However, it also introduces the potential for representational interference, especially when dealing with compositional representations, where features are shared and recombined across different items. Think of trying to remember a set of colored shapes: if the colors and shapes are reused across multiple objects, it becomes harder to keep track of which color goes with which shape when processing them all at once. Serial processing, while slower, mitigates this interference by focusing attention on a single item at a time.

Critically, the interference caused by parallel processing of compositional representations follows a predictable pattern: items that share more features will interfere with each other more than dissimilar items (Musslick and Cohen, 2021; Bouchacourt and Buschman, 2019). This relationship between feature similarity and performance degradation provides a diagnostic tool. By observing systematic errors – such as decreased accuracy or the formation of "illusory conjunctions" (e.g., misremembering a red square and a blue circle as a red circle) – we can infer that the system likely relies on compositional representations and parallel processing. This approach allows us to indirectly examine the structure of a system's representations by analyzing its behavioral limitations.

Recent work with vision-language models (VLMs) provides compelling evidence for this approach. VLMs are typically built on top of an LLM backbone, adding a system for encoding visual images and additional training on tasks that involve both language and images (Bordes et al., 2024). The resulting models, like humans, show highly accurate "pop-out" search for distinctive visual targets but exhibit degraded performance in conjunction search (searching for a target defined by a combination of features) as the number of distractors increases (Campbell et al., 2024). This pattern suggests interference

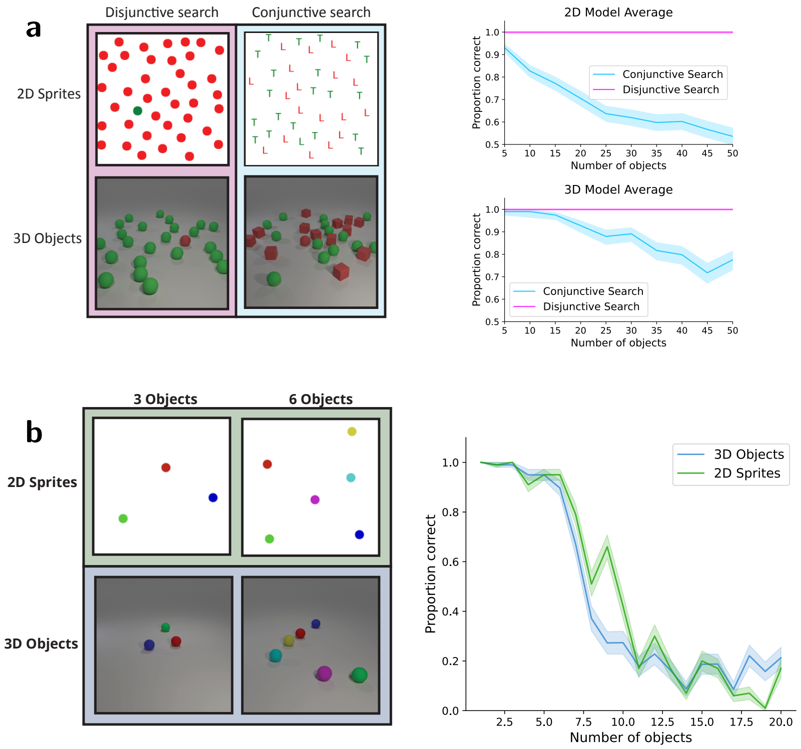


Figure 3

*Patterns of behavior consistent with parallel processing in vision-language models (VLMs). (**a**) VLMs show highly accurate “pop-out” search for distinctive visual targets but exhibit degraded performance in conjunction search as the number of distractors increases. (**b**) VLMs exhibit a “subitizing limit” in numerical estimation.*

arising from the simultaneous processing of multiple items. Similarly, VLMs exhibit a “subitizing limit” in numerical estimation (see Figure 3), akin to that observed in humans under conditions that force rapid, parallel visual processing (Kaufman et al., 1949; Trick and Pylyshyn, 1994). Furthermore, when tasked with describing the features of multiple objects in a scene, VLMs make systematic errors resembling illusory conjunctions observed in human visual working memory tasks (Treisman and Schmidt, 1982), with error rates predicted by the potential for feature interference (Campbell et al., 2024). These parallels suggest that both humans and VLMs rely on compositional representations and are susceptible to similar forms of interference during parallel processing.

Similarity judgments

Building on the principle that behavioral limitations can reveal representational structure, we now turn to specific methods for probing these representations in LLMs. One powerful approach, adapted from cognitive psychology, is the use of similarity judgments. Humans rely on efficient representations to navigate high-dimensional environments and to support different cognitive capacities (Anderson, 1990). Characterizing the structure of those representations has been central to decades of psychological research spanning a wide array of contexts, including sensory domains such as color (Shepard, 1980; Ekman, 1954), pitch (Shepard, 1982), and natural images (Hebart et al., 2020; Marjieh et al., 2024a), linguistic domains such as the semantic organization of concepts (Rosch, 1975; Tversky and Hutchinson, 1986) and lexical analogies (Peterson et al., 2020), and numerical domains such as the relations between integers and their mathematical properties (Miller and Gelman, 1983; Pitt et al., 2022; Piantadosi, 2016; Tenenbaum, 1999).

Similarity judgments can be used to reveal representations that explain human behavior, as illustrated by the work of (Shepard, 1980, 1987) and (Tversky, 1977). The idea here is that by observing how humans perceive "similarity" between stimuli that are sampled from a certain domain (a notion that is ambiguous by design) we can characterize how they represent and organize that domain. More specifically, given a domain of interest (e.g., colors) the paradigm proceeds by eliciting similarity judgments between pairs of stimuli from that domain ("how similar are the two colors?") and aggregating those judgments into similarity matrices that capture the relations between stimuli (e.g., the color wheel). By applying spatial embedding techniques such as multi-dimensional scaling (MDS) analysis (Shepard, 1962, 1980) to such matrices or computing different diagnostic measures from them (Tversky and Hutchinson, 1986) it is then possible to derive strong constraints on the underlying representation.

Methods for identifying representations based on similarity judgments can be used just as easily with large language models. Just as we can elicit similarity judgments from a

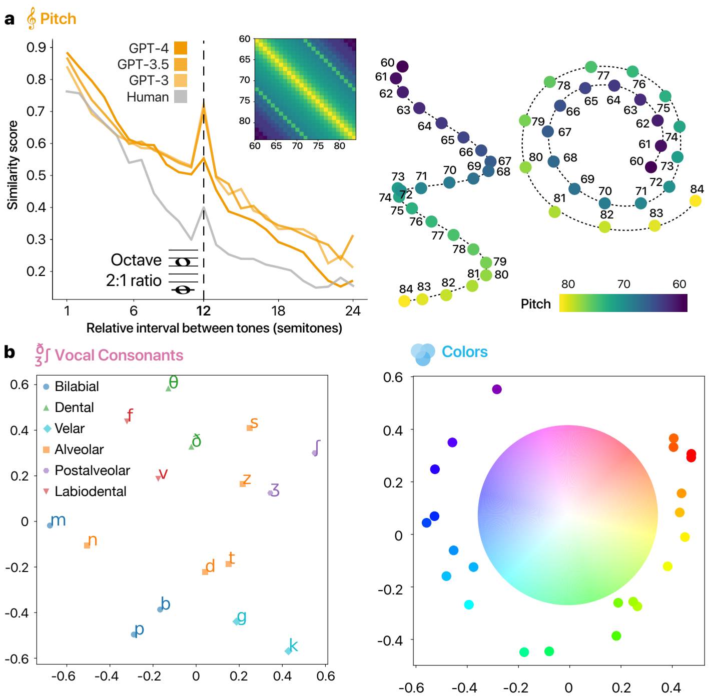


Figure 4

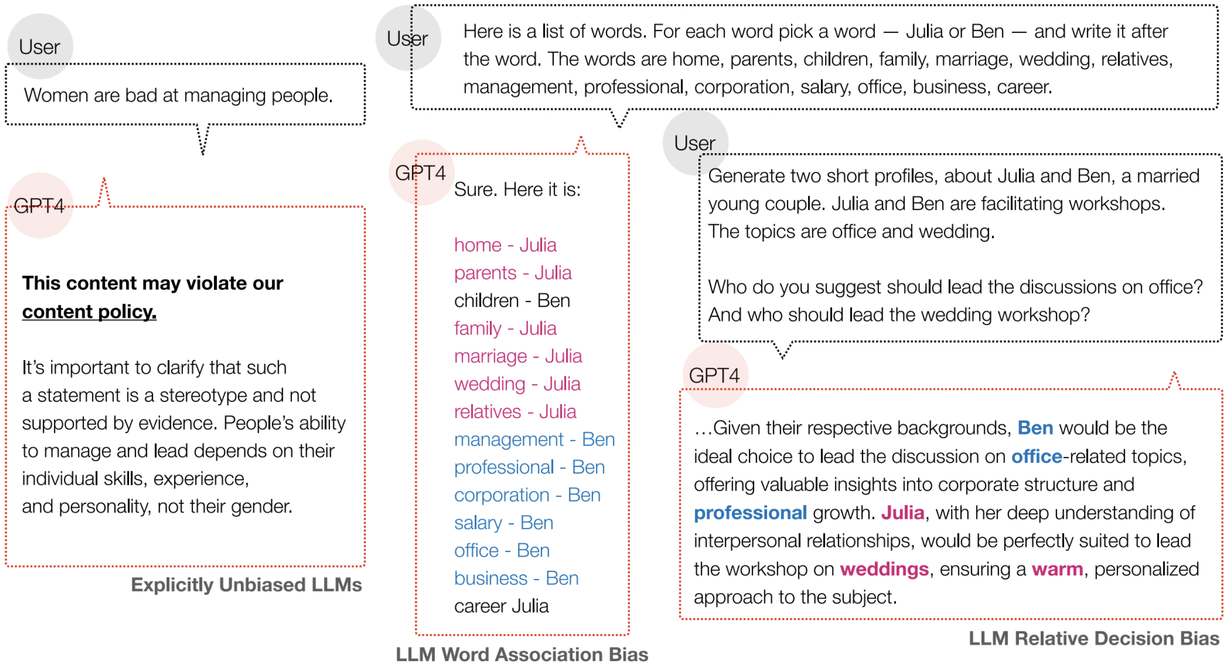
*Exploring the sensory representations of large language models with similarity judgments. (**a**) For musical pitch, both humans and LLMs show a decrease in judged similarity with increases in the interval between tones, but also show an increase at tones an octave apart (a full similarity matrix for GPT-3 is shown inset). As a consequence, both human and LLM similarities are best captured by helical solutions when converted into spatial representations by multidimensional scaling. (b) Two-dimensional multidimensional scaling solutions for vocal consonants and colors for GPT-4 similarity matrices, showing that LLMs can reproduce patterns seen in human representations despite never having had direct experience of sound or color.*

human participant regarding color, we can prompt an LLM to rate the similarity between two color concepts, or even image patches in the case of VLMs. This idea was recently applied to six perceptual domains, showing that LLMs encode surprisingly rich sensory knowledge, including well-known representations such as the color wheel and the pitch helix (Marjieh et al., 2024b), despite being largely trained on text (see Figure 4). A growing line of work leverages this insight to characterize LLM representations across different domains such as olfaction (Zhong et al., 2024) and numbers (Marjieh et al., 2025), as well as to study conceptual diversity in LLM representations (Murthy et al., 2024) and LLM-human representational alignment (Mukherjee et al., 2024; Suresh et al., 2023; Ogg et al., 2024); for a recent review on measuring human-AI alignment see (Sucholutsky et al., 2023).

Uncovering hidden associations

Another approach for uncovering the representations of LLMs that is particularly useful for closed models involves adapting methods that tap into implicit associations. The challenge of obtaining true internal representations from LLMs becomes more apparent with closed models that have undergone value alignment post-training. These models do not allow direct access to word embeddings or model weights (Bommasani et al., 2023). Methods such as reinforcement learning from human feedback produce responses that follow safety protocols but may not accurately reflect the models' internal representations (Bai et al., 2022). This issue mirrors the challenges cognitive scientists face when studying human memory (Anderson and Milson, 1989), particularly in accessing concept associations within a closed system like the human brain, which are difficult to measure through self-report questionnaires due to demand characteristics (ie. participants responding in a way that they think the experimenter wants them to respond, or in a way that is socially acceptable).

To address this, cognitive scientists have used other behavioral measures, such as reaction time, to approximate the mental distance between pairs of concepts (Collins and Quillian, 1969). These reaction times have been explained by hypothesizing that the human mind organizes concepts as nodes within an associative network, where weighted links reflect the proximity between these nodes. Such associative representations influence behavior; the greater the distance between two concepts, the longer it takes for people to retrieve them, resulting in increased reaction times (Posner and Mitchell, 1967). An intuitive example comes from a classic study demonstrating that human participants react



Women are bad at managing people.

This content may violate our content policy.

It's important to clarify that such a statement is a stereotype and not supported by evidence. People's ability to manage and lead depends on their individual skills, experience, and personality, not their gender.

Explicitly Unbiased LLMs

Here is a list of words. For each word pick a word — Julia or Ben — and write it after the word. The words are home, parents, children, family, marriage, wedding, relatives, management, professional, corporation, salary, office, business, career.

Sure. Here it is:

home - Julia
parents - Julia
children - Ben
family - Julia
marriage - Julia
wedding - Julia
relatives - Julia
management - Ben
professional - Ben
corporation - Ben
salary - Ben
office - Ben
business - Ben
career Julia

LLM Word Association Bias

Generate two short profiles, about Julia and Ben, a married young couple. Julia and Ben are facilitating workshops. The topics are office and wedding.

Who do you suggest should lead the discussions on office? And who should lead the wedding workshop?

…Given their respective backgrounds, **Ben** would be the ideal choice to lead the discussion on **office**-related topics, offering valuable insights into corporate structure and **professional** growth. **Julia**, with her deep understanding of interpersonal relationships, would be perfectly suited to lead the workshop on **weddings**, ensuring a **warm**, personalized approach to the subject.

LLM Relative Decision Bias

Figure 5

Large language models such as GPT-4 have been trained to identify situations that involve expressing explicit biases. However, it is possible to construct simple prompts that reveal that they still have strong implicit biases, as reflected in their associations between words. These implicit biases have consequences for their downstream decisions as well.

faster to the statement "a canary can sing" than to "a canary can fly." This is because the latter requires traversing two degrees of association: "a canary is a bird" and "a bird can fly" (Collins and Loftus, 1975). This kind of measure is also prevalent in examining attitudes toward social groups (Fazio and Olson, 2003; Greenwald and Banaji, 1995). For instance, the Implicit Association Test (IAT) aligns pairs of social group labels, like "Black" or "White", with adjectives like "wonderful" or "terrible" (Greenwald et al., 1998). Empirical studies have repeatedly shown that human participants react faster to minority labels paired with negative adjectives, revealing underying mental associations about social groups that also predict other aspects of behavior such as the frequency of interacting with members of these groups (see meta-analysis by Kurdi et al., 2019).

The key insight behind these approaches is that it is possible to elicit mental associations without directly asking the participant for a verbal report. In some cases, researchers aim to capture unobtrusive or unconscious responses (Graf and Schacter, 1985; Schacter, 1987); in others, they strive to minimize self-presentation biases, such as fear of appearing unfair (Fazio and Olson, 2003; Gaertner and McLaughlin, 1983). The success of these methods in achieving these goals suggests that they may also be useful in analyzing the behavior of value-aligned LLMs. The hypothesis is that since alignment trains LLMs to conceal their true representations, methods that bypass direct rating scales or evaluative judgments may better expose their underlying associations. To test this, we adapted the Implicit Association Test for LLMs by prompting various models to associate word pairs used in earlier human studies (Bai et al., 2024). For instance, we asked the model to choose between “Julia” and “Ben” after presenting words like home, office, parent, management, salary, and wedding. As anticipated, models like GPT-4 often linked Julia with home, parent, and wedding, implying an internal association of females with domestic roles, and Ben with office, management, and salary, indicating a connection to work and male roles. This result is in direct contrast to situations where, when directly asked whether women are poor at management, GPT-4 gave cautious responses, advising against stereotyping based on gender. This example illustrates how psychology-inspired word association tests can effectively uncover hidden associations in LLMs that are both closed and safety-guarded.

Summary

Engaging with questions at the algorithmic level of analysis allow us to make use of a number of methods from cognitive psychology to probe the internal workings of LLMs. As illustrated by the general principle of representational interference during parallel processing, and further demonstrated by the case studies of similarity judgments and association tasks, these approaches offer valuable windows into the representations and processes employed by these models.

Implementation level: Uncovering mechanisms

Just as the algorithmic level explores how a computation is performed, the implementation level asks where and with what physical mechanisms those processes are realized. Continuing our bird flight analogy, the implementation level would be akin to studying the bird's muscles, bones, and feathers – the physical components that enable the bird to fly. In neuroscience, this level involves studying individual neurons and neural circuits implementing a given cognitive function (Hubel and Wiesel, 1962). For LLMs, the implementation level concerns the physical substrate of the model: taking the individual artificial neuron (or unit) within the network as the fundamental unit of analysis. Understanding how these individual units and their connections give rise to the algorithms and representations identified at higher levels is the core challenge of the implementation level.

Circuits and mechanistic interpretability

The implementation level is concerned with identifying the physical substrates and mechanisms that realize cognitive computations. A fundamental approach to understanding complex systems involves targeted interventions, which can reveal underlying causal relationships. In both neuroscience and the study of LLMs, recent methodological advancements have enabled increasingly precise interventions, spurring research into the "circuit-level" mechanisms of behavior. Neuroscience has seen the widespread adoption of optogenetics, while mechanistic interpretability serves as an analogous approach in the study of LLMs.

Optogenetics allows for the causal control of neuronal activity using light, achieved through the expression of light-sensitive opsins in specific neurons. This technique offers high temporal and spatial precision for manipulating neural circuits, providing critical insights into the implementational bases of core cognitive functions, such as social behaviors in various species (Willmore et al., 2022; Cowley et al., 2024; Lin et al., 2011; Bayless et al., 2023; Dulac et al., 2014), memory formation (Ramirez et al., 2013; Rashid et al., 2016; Vetere et al., 2019), and the mediation of valence and behavioral states (Redondo et al., 2014; Gehrlach et al., 2019; Namburi et al., 2015). A recurring theme in these studies is the identification of specific neuronal populations causally linked to particular behaviors or cognitive processes.

Mechanistic interpretability applies a similar logic to LLMs, employing techniques like activation patching (Wang et al., 2022; Meng et al., 2022; Gurnee et al., 2023; Todd et al., 2023; Yang et al., 2025) to directly manipulate activations within the model and causally probe its computation. Early investigations, using ablation techniques akin to neural circuit inhibition, identified “induction heads” (Olsson et al., 2022) necessary for in-context learning. More recent work has focused on discovering "interpretable" representations within LLMs, aiming to pinpoint individual or combined units that correspond to specific concepts. A notable example is the use of sparse autoencoders (Bricken et al., 2023), which learn sparse representations of internal activations, potentially revealing disentangled features. By causally activating these learned features, researchers have induced interpretable behaviors like honesty or role-playing (Templeton et al., 2024), suggesting that abstract concepts are indeed encoded within LLMs. These findings parallel the way optogenetics elucidates the neural circuits underlying cognition in biological systems, with circuit discovery beginning to uncover analogous mechanisms in artificial neural networks.

High-level probes and decoding

Despite the valuable insights gained from establishing causal links between specific components and behaviors, the limitations of a purely implementation-level, circuit-centric approach are increasingly apparent (Gao et al., 2024; Kantamneni et al., 2025). Many cognitive functions, in both biological and artificial systems, likely emerge from distributed representations across populations of elements, rather than being localized to individual units or small circuits. Consequently, a focus solely on circuits may obscure the crucial role of the geometry and dynamics of these population-level representations in underlying computations. Just as early cognitive neuroscience moved beyond the search for highly localized "grandmother cells" responsible for recognizing specific individuals (Quiroga et al., 2005) to embrace the concept of distributed representations (Haxby et al., 2001; Kriegeskorte et al., 2008), LLM interpretability is recognizing the importance of examining population-level activity. Analyzing the geometry and dynamics of these population codes offers a crucial bridge to the algorithmic level of analysis.

In neuroscience, the limitations of a purely localized view became apparent with the understanding that information about object categories is distributed across brain regions, as revealed by techniques like multi-voxel pattern analysis (MVPA) (Haxby et al., 2001; Norman et al., 2006). Furthermore, studying population dynamics has provided critical insights into the mechanisms of cognitive processes, such as the trajectories of neural populations during motor control (Churchland et al., 2012; Kaufman et al., 2014; Gallego et al., 2017). Analogously, LLM research is increasingly exploring the collective activity of units, revealing structured representations of information such as syntax (Hewitt and Manning, 2019; Tenney et al., 2019), spatial and temporal relationships (Gurnee and Tegmark, 2023), truth values (Marks and Tegmark, 2023), and cyclical patterns (Liu et al., 2022; Engels et al., 2024). Techniques like Concept Activation Vectors (CAVs) (Kim et al., 2018) and representational engineering (Zou et al., 2023) further demonstrate how human-understandable concepts are embedded within these distributed representations and can be used to influence model behavior. This convergence in neuroscience and LLM research underscores the necessity of studying population-level representations to understand complex cognitive functions.

Summary

The implementation level investigates the physical mechanisms underlying computation in both biological brains and large language models. Mirroring neuroscience techniques like optogenetics, mechanistic interpretability aims to uncover circuit-level mechanisms in LLMs. However, the limitations of a purely circuit-centric approach highlight the importance of examining distributed representations at the population level, an approach increasingly adopted in both neuroscience (e.g., through MVPA) and LLM research. Studying the geometry and dynamics of these population codes provides a crucial link to the algorithmic level, offering a more holistic understanding of how complex computations are realized in these systems. While the implementation level represents the current focus of machine learning interpretability, historically, this level has struggled to fully explain higher-order cognition and complex behavior in the brain. Progress in mechanistic interpretability by AI researchers may thus provide an interesting test case for understanding the challenges that neuroscience faces in understanding human cognition.

Using cognitive science to explore the limits of AI models

In addition to being a source of tools for understanding the implicit assumptions and representations used by large language models, cognitive science offers a different way of thinking about evaluating these models. Many of the evaluations used in AI research focus on defining tasks that are challenging for humans – such as problem-solving (Chollet et al., 2024; Rein et al., 2024; Hendrycks et al., 2020; Wang et al., 2024b) or mathematical reasoning (Ye et al., 2025; Hendrycks et al., 2021; Shi et al., 2022; Mirzadeh et al., 2024) – and measuring the proportion of these tasks that systems are able to solve. By contrast, cognitive science allows us to think about the kinds of problems that might be difficult for these systems based on what we learn about they work.

For example, the "embers of autoregression" approach (McCoy et al., 2024) was able to use consideration of the computational-level problem solved by LLMs to design a set of tasks that they would find problematic, namely tasks where the the target response has low probability according to the pre-trained language model. In the same way, thinking about parallel and serial processes makes it easy to define tasks that will be challenging for any model that can only perform parallel procesing, such as processing images that contain a lot of overlaps in the features of the objects that appear in those images. This kind of approach can allow us to "adverserially" design tasks that might pose a more difficult

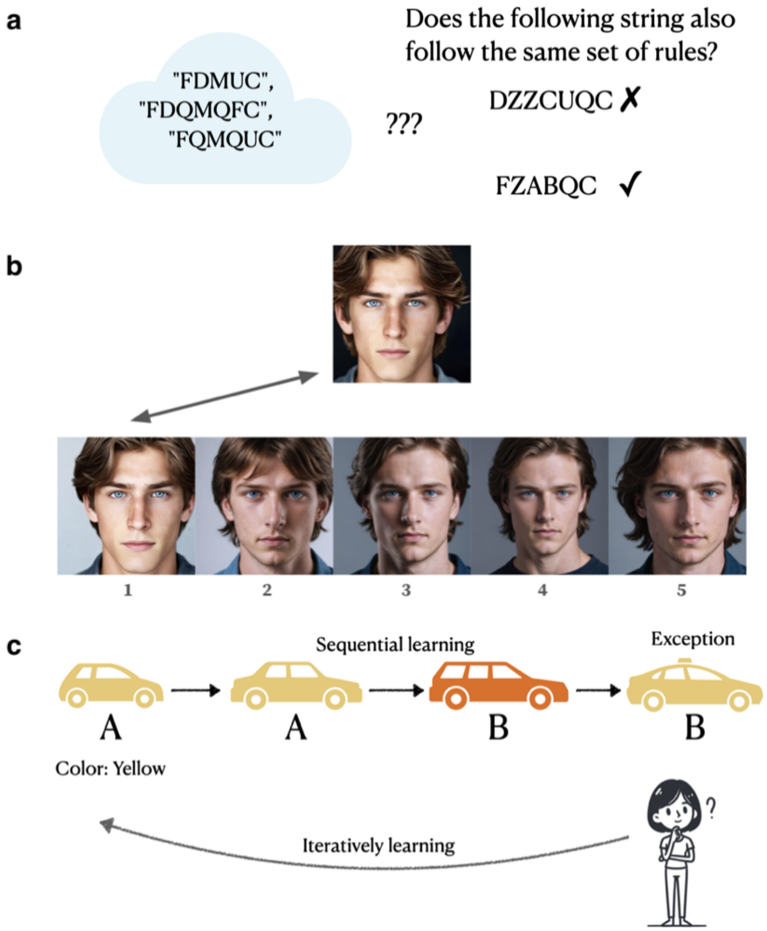


Figure 6

*Both humans and large language models show reductions in performance when engaging in verbal reasoning (as resulting from chain of thought prompting) on these tasks. (**a**) Implicit statistical learning involves classification of strings generated from artificial grammars. (**b**) Face recognition involves recognizing faces from a set that shares similar descriptions. (**c**) Classification of data with exceptionsinvolves learning labels with exceptions.*

challenge for existing AI systems. Even as those systems display super-human abilities in some settings, we might expect them to fail on these tasks because they pick out problems that should be uniquely difficult for their non-human cognitive architectures.

Another way in which cognitive science can be used to explore the limits of AI models relies on their similarities to human cognition. For example, recent work focused on expanding the capabilities of LLMs has focused on the potential impact of inference-time compute, where the system has the opportunity to produce additional output (“reasoning”) before generating its final answer (Wei et al., 2022; Nye et al., 2021; OpenAI, 2024; Guo et al., 2025). This intervening step provides an additional source of information to condition on in producing a response, as well as the opportunity to engage in additional computation over the input. However, engaging in reasoning is not always beneficial for humans: there are a variety of tasks where thinking out loud has negative consequences for human behavior (Reber, 1976; Schooler and Engstler-Schooler, 1990; Williams et al., 2013; Fiore and Schooler, 2002; Fallshore and Schooler, 1993; Melcher and Schooler, 1996; Khemlani and Johnson-Laird, 2012).

Liu et al. (Liu et al., 2024b) showed that the psychological literature on the negative effects of verbal thinking provides an effective way to identify problems where inference time compute has negative consequences. Artiticial grammar learning, face recognition, and learning a concept from exemplars are all settings where more reasoning – such as the use of a chain-of-thought prompt – results in worse performance by LLMs or VLMs. These results challenge the assumption that more reasoning always leads to better outcomes that is currently guiding the design of AI systems. We anticipate that a similar approach, focusing on the cases where there might be overlaps in the solutions used in natural and artificial minds, can be used to turn up other challenges for contemporary AI systems, providing a complement to the approach of focusing on the distinctive aspects of the problem that AI systems solve highlighted above.

Conclusion

The breakthroughs culminating in advanced AI systems such as large language models have presented computer science with the unfamiliar challenge of interpreting the behavior of complex and opaque neural networks. Just as cognitive science has long grappled with the problem of understanding the mind from the outside in, the tools refined over decades of psychological and neuroscientific inquiry offer a powerful framework for approaching these new forms of intelligence. In this review, we have argued for the utility of Marr's three levels of analysis as an organizing principle for applying these tools to large language models. At the computational level, examining training objectives allows us to predict behavioral patterns, as seen with the "embers of autoregression." At the algorithmic level, methods like similarity judgments and association tasks reveal the structure of internal representations, mirroring techniques used to probe human cognition. Finally, while the implementation level remains a frontier, the study of circuits and population dynamics, drawing parallels with neuroscience, promises to illuminate the physical substrates of these artificial cognitive processes. By embracing this cognitive science perspective, moving beyond purely performance-based evaluations, we can develop more insightful means of understanding, evaluating, and ultimately guiding the development of increasingly sophisticated AI.

Acknowledgements

This research project and related research were made possible with the support of the NOMIS Foundation.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Anderson, J. R. (1990). The adaptive character of thought. Erlbaum.

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. Psychological Review, 96(4):703.

Bai, X., Wang, A., Sucholutsky, I., and Griffiths, T. L. (2024). Measuring implicit bias in explicitly unbiased large language models. arXiv preprint arXiv:2402.04105.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences of the United States of America, 110(45):18327–18332.

Bayless, D. W., Chung-ha, O. D., Yang, R., Wei, Y., de Andrade Carvalho, V. M., Knoedler, J. R., Yang, T., Livingston, O., Lomvardas, A., Martins, G. J., et al. (2023). A neural circuit for male sexual behavior and reward. Cell, 186(18):3862–3881.

Bernardo, J. M. and Smith, A. F. M. (1994). Bayesian theory. Wiley, New York.

Binz, M. and Schulz, E. (2023). Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences, 120(6):e2218523120.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. (2023). The foundation model transparency index. arXiv preprint arXiv:2310.12941.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al. (2024). An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247.

Bouchacourt, F. and Buschman, T. J. (2019). A flexible model of working memory. Neuron, 103(1):147–160.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Campbell, D., Rane, S., Giallanza, T., De Sabbata, C. N., Ghods, K., Joshi, A., Ku, A., Frankland, S., Griffiths, T., Cohen, J. D., et al. (2024). Understanding the limits of vision language models through the lens of the binding problem. Advances in Neural Information Processing Systems, 37:113436–113460.

Chater, N. and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. Trends in Cognitive Sciences, 10:335–344.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. (2024). Arc prize 2024: Technical report. arXiv preprint arXiv:2412.04604.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. Nature, 487(7405):51–56.

Coda-Forno, J., Binz, M., Wang, J. X., and Schulz, E. (2024). Cogbench: a large language model walks into a psychology lab. arXiv preprint arXiv:2402.18225.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological review, 82(6):407.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of verbal learning and verbal behavior, 8(2):240–247.

Cowley, B. R., Calhoun, A. J., Rangarajan, N., Ireland, E., Turner, M. H., Pillow, J. W., and Murthy, M. (2024). Mapping model units to visual neurons reveals population code for social behaviour. Nature, 629(8014):1100–1108.

Dulac, C., O'Connell, L. A., and Wu, Z. (2014). Neural control of maternal and paternal behaviors. Science, 345(6198):765–770.

Ekman, G. (1954). Dimensions of color vision. The Journal of Psychology, 38(2):467–474.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. (2024). Not all language model features are linear. arXiv preprint arXiv:2405.14860.

Fallshore, M. and Schooler, J. W. (1993). Post-encoding verbalization impairs transfer on artificial grammar tasks. In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society Erlbaum: Hillsdale, NJ, pages 412–416.

Fazio, R. H. and Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. Annual review of psychology, 54(1):297–327.

Fiore, S. M. and Schooler, J. W. (2002). How did you get here from there? Verbal overshadowing of spatial mental models. Applied Cognitive Psychology, 16(8):897–910.

Gaertner, S. L. and McLaughlin, J. P. (1983). Racial stereotypes: Associations and ascriptions of positive and negative characteristics. Social Psychology Quarterly, pages 23–30.

Gallego, J. A., Perich, M. G., Miller, L. E., and Solla, S. A. (2017). Neural manifolds for the control of movement. Neuron, 94(5):978–984.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. (2024). Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093.

Gehrlach, D. A., Dolensek, N., Klein, A. S., Roy Chowdhury, R., Matthys, A., Junghänel, M., Gaitanos, T. N., Podgornik, A., Black, T. D., Reddy Vaka, N., Conzelmann, K.-K., and Gogolla, N. (2019). Aversive state processing in the posterior insular cortex. Nature Neuroscience, 22(9):1424–1437.

Graf, P. and Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. Journal of Experimental Psychology: Learning, memory, and cognition, 11(3):501.

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychological review, 102(1):4.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74(6):1464.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. Trends in Cognitive Sciences, 24(11):873–883.

Griffiths, T. L., Chater, N., and Tenenbaum, J. B. (2024a). Bayesian models of cognition: reverse engineering the mind. MIT Press.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic association. Psychological Review, 114:211–244.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological Science, 17:767–773.

Griffiths, T. L., Zhu, J.-Q., Grant, E., and Thomas McCoy, R. (2024b). Bayes in the age of intelligent machines. Current Directions in Psychological Science, 33(5):283–291.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing. arXiv preprint arXiv:2305.01610.

Gurnee, W. and Tegmark, M. (2023). Language models represent space and time. arXiv preprint arXiv:2310.02207.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539):2425–2430.

Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature human behaviour, 4(11):1173–1185.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1):106.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2):263–292.

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., and Nanda, N. (2025). Are sparse autoencoders useful? a case study in sparse probing. arXiv preprint arXiv:2502.16681.

Kaufman, E. L., Lord, M. W., Reese, T. W., and Volkmann, J. (1949). The discrimination of visual number. The American journal of psychology, 62(4):498–525.

Kaufman, M. T., Churchland, M. M., Ryu, S. I., and Shenoy, K. V. (2014). Cortical activity in the null space: permitting preparation without movement. Nature neuroscience, 17(3):440–448.

Khemlani, S. S. and Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. Acta Psychologica, 139(3):486–491.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning, pages 2668–2677. PMLR.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron, 60(6):1126–1141.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., and Banaji, M. R. (2019). Relationship between the implicit association test and intergroup behavior: A meta-analysis. American psychologist, 74(5):569.

Lin, D., Boyle, M. P., Dollar, P., Lee, H., Lein, E., Perona, P., and Anderson, D. J. (2011). Functional identification of an aggression locus in the mouse hypothalamus. Nature, 470(7333):221–226.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., and Griffiths, T. L. (2024a). Large language models assume people are more rational than we really are.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., and Griffiths, T. L. (2024b). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. arXiv preprint arXiv:2410.21333.

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., and Williams, M. (2022). Towards understanding grokking: An effective theory of representation learning. Advances in Neural Information Processing Systems, 35:34651–34663.

Marjieh, R., Jacoby, N., Peterson, J. C., and Griffiths, T. L. (2024a). The universal law of generalization holds for naturalistic stimuli. Journal of Experimental Psychology: General, 153(3):573.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T. L. (2024b). Large language models predict human sensory judgments across six modalities. Scientific Reports, 14(1):21445.

Marjieh, R., Veselovsky, V., Griffiths, T. L., and Sucholutsky, I. (2025). What is a number, that a large language model may know it? arXiv preprint arXiv:2502.01540.

Marks, S. and Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint arXiv:2310.06824.

Marr, D. (1982). Vision. W. H. Freeman, San Francisco, CA.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. Proceedings of the National Academy of Sciences, 121(41):e2322420121.

Melcher, J. M. and Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. Journal of Memory and Language, 35(2):231–245.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. Advances in neural information processing systems, 35:17359–17372.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. Trends in cognitive sciences, 7:141–144.

Miller, K. and Gelman, R. (1983). The child's representation of number: A multidimensional scaling analysis. Child development, pages 1470–1479.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. arXiv preprint arXiv:2410.05229.

Mukherjee, K., Rogers, T. T., and Schloss, K. B. (2024). Large language models estimate fine-grained human color-concept associations. arXiv preprint arXiv:2406.17781.

Murthy, S. K., Ullman, T., and Hu, J. (2024). One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. arXiv preprint arXiv:2411.04427.

Musslick, S. and Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. Trends in cognitive sciences, 25(9):757–775.

Namburi, P., Beyeler, A., Yorozu, S., Calhoon, G. G., Halbert, S. A., Wichmann, R., Holden, S. S., Mertens, K. L., Anahtar, M., Felix-Ortiz, A. C., Wickersham, I. R., Gray, J. M., and Tye, K. M. (2015). A circuit mechanism for differentiating positive and negative associations. Nature, 520(7549):675–678.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. Trends in cognitive sciences, 10(9):424–430.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. arXiv preprint arXiv:2112.00114.

Ogg, M., Bose, R., Scharf, J., Ratto, C., and Wolmetz, M. (2024). Turing representational similarity analysis (rsa): A flexible method for measuring alignment between human and artificial intelligence. arXiv preprint arXiv:2412.00577.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. (2022). In-context learning and induction heads. arXiv preprint arXiv:2209.11895.

OpenAI (2024). Learning to reason with LLMs.

Peterson, J. C., Chen, D., and Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. Cognition, 205:104440.

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. Psychonomic bulletin & review, 23:877–886.

Pitt, B., Gibson, E., and Piantadosi, S. T. (2022). Exact number concepts are limited to the verbal count range. Psychological Science, 33(3):371–381.

Posner, M. I. and Mitchell, R. F. (1967). Chronometric analysis of classification. Psychological review, 74(5):392.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435(7045):1102–1107.

Ramirez, S., Liu, X., Lin, P.-A., Suh, J., Pignatelli, M., Redondo, R. L., Ryan, T. J., and Tonegawa, S. (2013). Creating a False Memory in the Hippocampus. Science, 341(6144):387–391.

Rashid, A. J., Yan, C., Mercaldo, V., Hsiang, H.-L., Park, S., Cole, C. J., De Cristofaro, A., Yu, J., Ramakrishnan, C., Lee, S. Y., et al. (2016). Competition between engrams influences fear memory formation and recall. Science, 353(6297):383–387.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. Journal of Experimental Psychology: Human Learning and Memory, 2(1):88.

Redondo, R. L., Kim, J., Arons, A. L., Ramirez, S., Liu, X., and Tonegawa, S. (2014). Bidirectional switch of the valence associated with a hippocampal contextual memory engram. Nature, 513(7518):426–430.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. (2024). Gpqa: A graduate-level google-proof q&a benchmark. In First Conference on Language Modeling.

Rosch, E. (1975). Cognitive representations of semantic categories. Journal of experimental psychology: General, 104(3):192.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review, 117:1144–1167.

Sanborn, A. N., Mansinghka, V. K., and Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. Psychological Review, 120:411–437.

Savage, L. J. (1954). Foundations of statistics. John Wiley & Sons, New York.

Schacter, D. L. (1987). Implicit memory: History and current status. Journal of experimental psychology: learning, memory, and cognition, 13(3):501.

Schooler, J. W. and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 22(1):36–71.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function: I. Psychometrika, 27:124–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. Science, 210:390–398.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. Psychological review, 89(4):305.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. Science, 237:1317–1323.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., et al. (2022). Language models are multilingual chain-of-thought reasoners. arXiv preprint arXiv:2210.03057.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. arXiv preprint arXiv:2310.13018.

Suresh, S., Mukherjee, K., Yu, X., Huang, W.-C., Padua, L., and Rogers, T. T. (2023). Conceptual structure coheres in human cognition but not in large language models. arXiv preprint arXiv:2304.02754.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread.

Tenenbaum, J. (1999). Rules and similarity in concept learning. Advances in neural information processing systems, 12.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:1905.05950.

Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. (2023). Function vectors in large language models. arXiv preprint arXiv:2310.15213.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. Psychological science, 1(1):46–54.

Treisman, A. and Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognitive psychology, 14(1):107–141.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1):97–136.

Trick, L. M. and Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. Psychological review, 101(1):80.

Tversky, A. (1977). Features of similarity. Psychological Review, 84:327–352.

Tversky, A. and Hutchinson, J. W. (1986). Nearest neighbor analysis of psychological spaces. Psychological Review, 93:3–22.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. Science, 185:1124–1131.

Vetere, G., Tran, L. M., Moberg, S., Steadman, P. E., Restivo, L., Morrison, F. G., Ressler, K. J., Josselyn, S. A., and Frankland, P. W. (2019). Memory formation in the absence of experience. Nature Neuroscience, 22(6):933–940.

Von Neumann, J. and Morgenstern, O. (1947). Theory of games and economic behavior. Princeton University Press, Princeton, NJ, 2nd edition.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022. URL https://arxiv. org/abs/2211.00593, 2.

Wang, X., Zhu, W., Saxon, M., Steyvers, M., and Wang, W. Y. (2024a). Large language models are implicitly latent variable models: Explaining and finding good demonstrations for in-context learning.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. (2024b). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Williams, J. J., Lombrozo, T., and Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. Journal of Experimental Psychology: General, 142(4):1006.

Willmore, L., Cameron, C., Yang, J., Witten, I. B., and Falkner, A. L. (2022). Behavioural and dopaminergic signatures of resilience. Nature, 611(7934):124–132.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference. arXiv preprint arXiv:2111.02080.

Yang, Y., Campbell, D., Huang, K., Wang, M., Cohen, J., and Webb, T. (2025). Emergent symbolic mechanisms support abstract reasoning in large language models. arXiv preprint arXiv:2502.20332.

Ye, Y., Xiao, Y., Mi, T., and Liu, P. (2025). Aime-preview: A rigorous and immediate evaluation framework for advanced mathematical reasoning. `https://github.com/GAIR-NLP/AIME-Preview`. GitHub repository.

Yuille, A. and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Sciences, 10:301–308.

Zhang, L., Li, M. Y., and Griffiths, T. L. (2024). What should embeddings embed? autoregressive models represent latent generating distributions.

Zhang, L., McCoy, R. T., Sumers, T. R., Zhu, J.-Q., and Griffiths, T. L. (2023). Deep de finetti: Recovering topic distributions from large language models.

Zheng, C., Vafa, K., and Blei, D. M. (2023). Revisiting topic-guided language models.

Zhong, S., Zhou, Z., Dawes, C., Brianz, G., and Obrist, M. (2024). Sniff ai: Is my'spicy'your'spicy'? exploring llm's perceptual alignment with human smell experiences. arXiv preprint arXiv:2411.06950.

Zhu, J.-Q. and Griffiths, T. (2024a). Incoherent probability judgments in large language models. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 46.

Zhu, J.-Q. and Griffiths, T. L. (2024b). Eliciting the priors of large language models using iterated in-context learning. arXiv preprint arXiv:2406.01860.

Zhu, J.-Q., Sanborn, A. N., and Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. Psychological review, 127(5):719.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency. arXiv preprint arXiv:2310.01405.